GREAT LAKES
INSTITUTE OF MANAGEMENT
Global Mindset - Indian Roots

GREAT LEARNING

PGP in Data Science and Engineering

Interim Report for:

Customer_Spending_Prediction

Customer_Spent_at_ClubMahindra_Resort

Group Members:	**Mentor:**
Bhargavi	Anjana Aggarwal
Umang Ojha	
Majid Azmi	
Meraj Ahmed	
Gaurav Jadhav	
Pawar Laxmi Devi	

Contents

Part 1 Project Background 3

Part 2 About the Data..3

Part 3 Missing Values Treatment..4

Part 4 Data Cleaning...4

4.1 **Steps taken to clean the data**...5

Part 5 Exploratory Data Analysis 5

5.1 Univariate Analysis for continuous features..5

5.2 Univariate Analysis for Categorical features..7

5.3 Bivariate Analysis for continuous features...10

5.4 Bivariate Analysis between Categorical features and Target variables..............................10

5.5 Checking if multicollinearity exists between continuous variables..................................13

5.6 Statistical Analysis..13

Part 6 Feature Engineering..14

Part 7 Feature Selection..16

Part 8 Model Building..16

8.1 First Approach..17

8.2 Second Approach..17

8.3 Third Approach..17

Part 9 Score after changing the nature of prediction from regression to classification..................18

Recommendation...18

Conclusion……...19

Part 1 Project Background

With the increase in tourism across the globe and due to availability of large data, travel companies wants to use this available data to ensure pricing and operational efficiency with the objectives of making profit by availing different   holiday packages to its customers along with high standard hospitality based on season and other different factors.

Main Objectives: To predict amount spent per room per night by customers so as to increase profit along with availing different holiday packages to the customers.

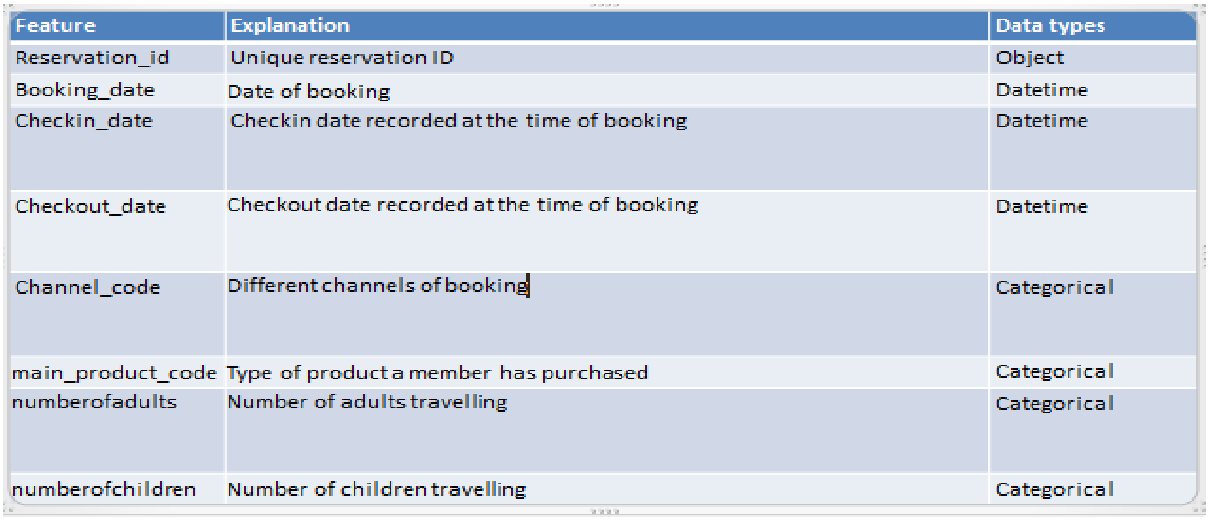
This project will enable the travel company (‘*Club Mahindra’*) to have more granular understanding of nature of customers spending for stay at resort and to know factors leading to demand of different types of resort, region based on the seasonality on the basis of which it can avail different offers and holiday packages to the customers .

By analysing Season of Holiday, Residence State of member ,state in which resort is located, Age bucket of the member and other features we can have insights regarding the demand of different resorts in different region, types of rooms being preferred by different age group, customers from states are tend to travel more and all the use fill insights to improve profit by operational efficiency.

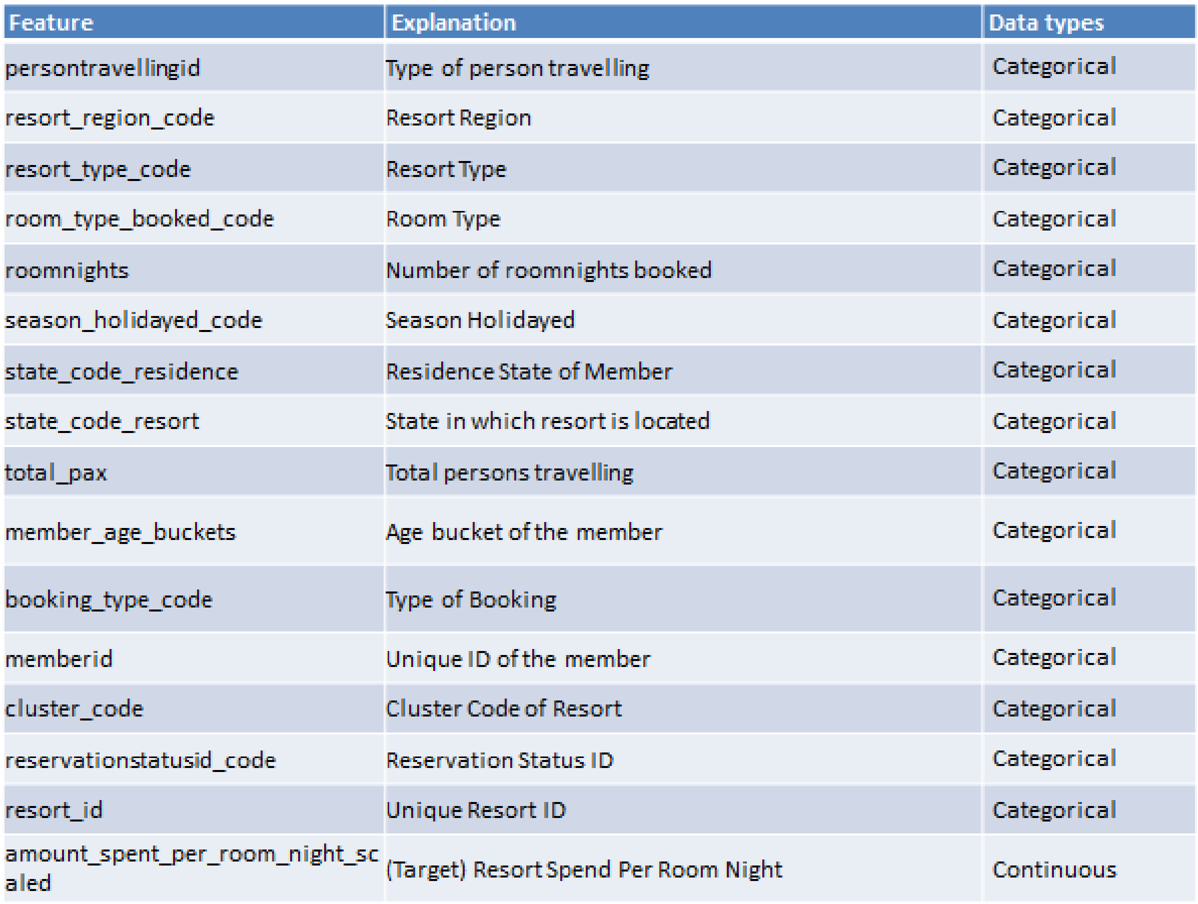
Part 2 About the Data

This data set consists of 3,41,424 observations, with 4,878 missing values. Each observation represents distinct customers with different Unique_id who booked the resorts for travelling purposes. For each observation, the dataset records 24 features that stand for both quantitative and qualitative attributes of customers.The data contains the information for the years 2015,2016,2017,2018.

There is single continuous output variable that denotes amount being paid by customers for each room per night. This feature is being scaled to hide the actual amount paid by the customers.



Feature	Explanation	Data types
Reservation_id	Unique reservation ID	Object
Booking_date	Date of booking	Datetime
Checkin_date	Checkin date recorded at the time of booking	Datetime
Checkout_date	Checkout date recorded at the time of booking	Datetime
Channel_code	Different channels of booking	Categorical
main_product_code	Type of product a member has purchased	Categorical
numberofadults	Number of adults travelling	Categorical
numberofchildren	Number of children travelling	Categorical



Feature	Explanation	Data types
persontravellingid	Type of person travelling	Categorical
resort_region_code	Resort Region	Categorical
resort_type_code	Resort Type	Categorical
room_type_booked_code	Room Type	Categorical
roomnights	Number of roomnights booked	Categorical
season_holidayed_code	Season Holidayed	Categorical
state_code_residence	Residence State of Member	Categorical
state_code_resort	State in which resort is located	Categorical
total_pax	Total persons travelling	Categorical
member_age_buckets	Age bucket of the member	Categorical
booking_type_code	Type of Booking	Categorical
memberid	Unique ID of the member	Categorical
cluster_code	Cluster Code of Resort	Categorical
reservationstatusid_code	Reservation Status ID	Categorical
resort_id	Unique Resort ID	Categorical
amount_spent_per_room_night_scaled	(Target) Resort Spend Per Room Night	Continuous

From the above details about the features we can see that except **booking_date, checkin_date, checkout_date** and **amout_spent_per_room_night_scaled** (target variable) all the variables are categorical in nature.

Out of total missing values,114 missing values are in **season_holidayed_code** and 4764 are in **state_code_residence**.

Part 3 Missing Values Treatment

Missing values were removed because they only consisted 1.41 % of all the observation. Therefore, replacing these missing values with mode was not going to have effect on our scores.

Part 4 Data Cleaning

Several changes were made to prepare the dataset for analysing.

While doing feature engineering using **booking_date, checkin_date, checkout_date to know the** exact number of days of stay. We found that in 73756 cases booking_date is greater than checkin_date(*practically booking_date and checkin_date can be same or checking date should be* *more than booking date*),in 45514 checking_date is more than checkout_date(*practically checkout_date and checkin_date can be same or check_in date should be less than checkout_date)* and 75835 cases booking date is more than checkout_date(*practically booking_date and checkout_date can be same or checkout_ date should be more than booking_ date).*

While checking for room_nights we found that there was one case where value was in negative case.

While for checking for total number of visitors by adding **numberofchildren** and **numberofadults** we found that 23 cases were found to be with 0 visitors but for these observations the rent was paid.

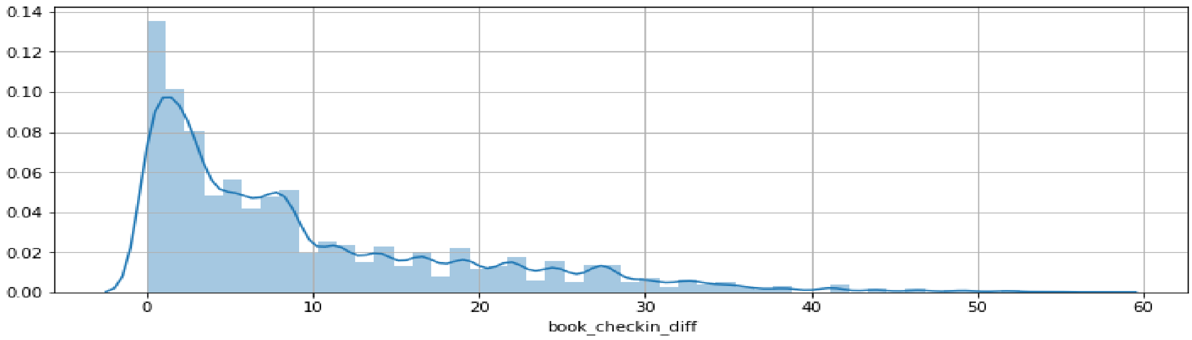
4.1 Steps taken to clean the data:

- In case of date deformities which have been mentioned above ,we removed all observation (total of 1,95,105) as we know that these deformities are not replaceable (dates cannot be replaced by our own rational wisdom) as we have to work on the data been given by source.
- In case of room_night we only had one negative value which we removed which doesn't effect our data.
- In case of total_number of people only 23 cases were there with 0 visitors therefore we removed those observations.

In total we removed around 1,95,129 rows with deformities which were either not replaceable of were very few in numbers.

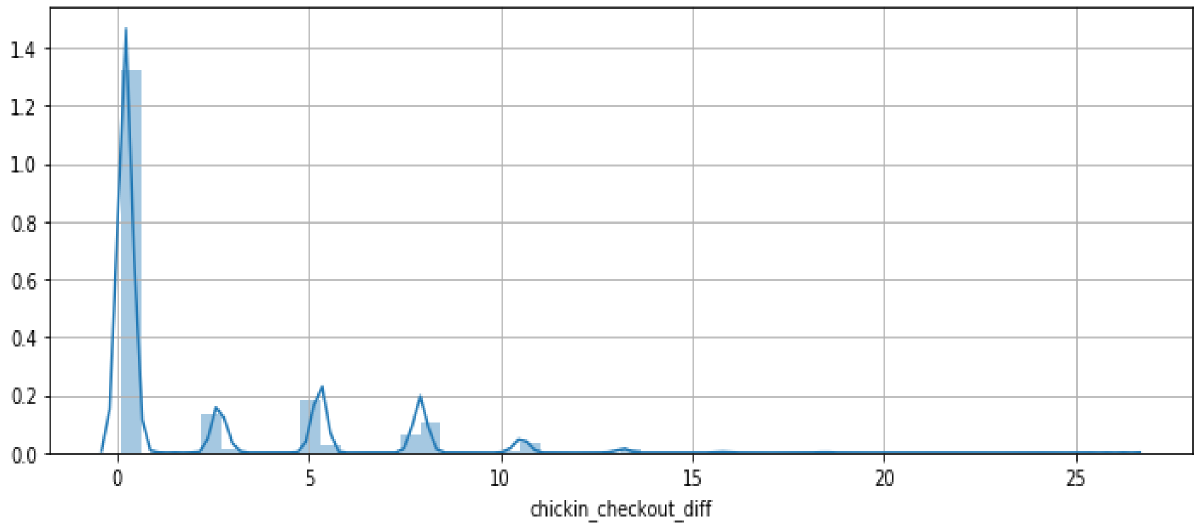
Part 5 Exploratory Data Analysis

5.1 Univariate Analysis for continuous features

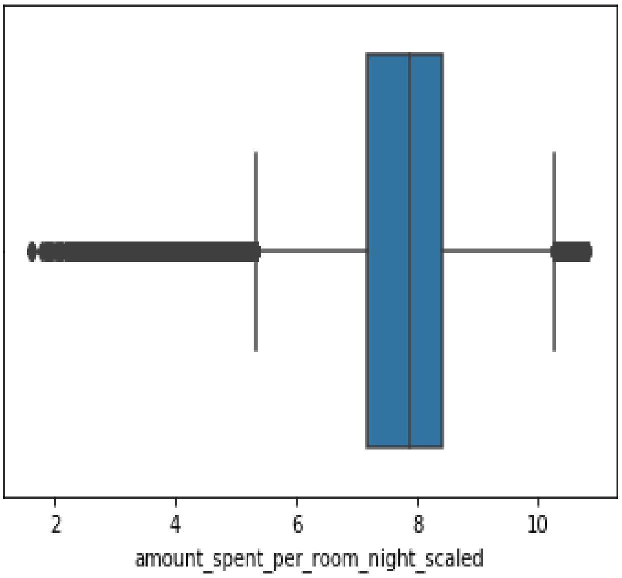


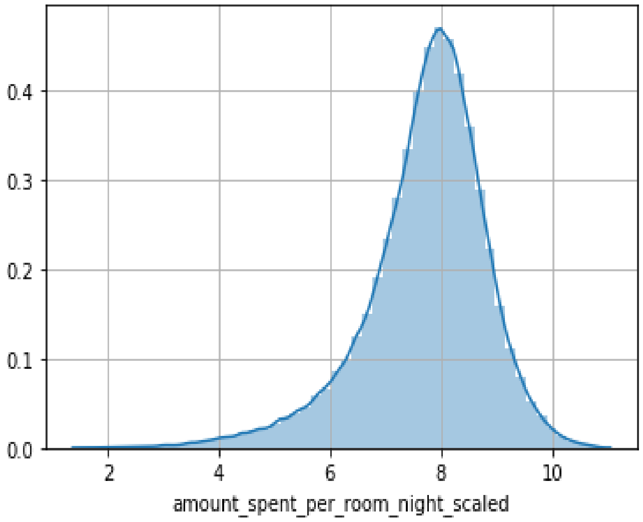
The above distribution plots show distribution of **book_checkin_diff** which is a feature engineered variable .This feature is created by subtracting booking_date from checkin_date to know the how early customers are booking the rooms before check_in .Using this feature we can analyse if our client is using flexi_fare system or if customers are getting any discount in rent for early bookings.

While analysing this distplot we can see that it is a left skewed plot which means that the majority of customers are booking the rooms within 5 days before checking in.



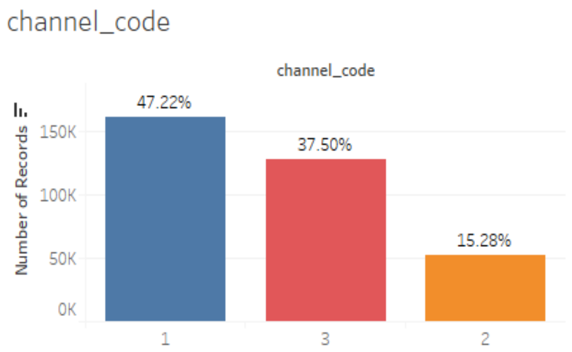
The above distribution plot shows distribution of checkin_checkout_diff which has been feature engineered by subtracting by checkin_date from checkout_date to know the how many days customers tend to stay in resort. After observing the above distplot we can say that there are 5 groups of customers which stay for different number of days. But most of the customers either stays for 1 day or 2 days.

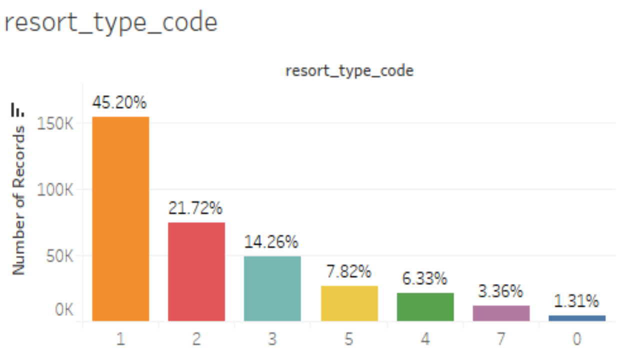


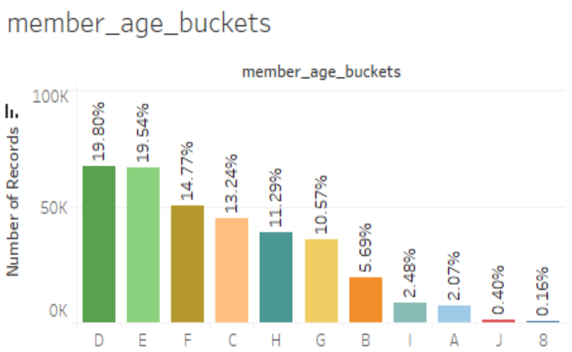


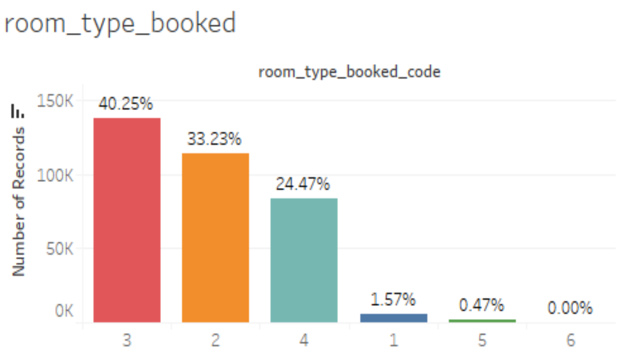
The above distribution plot is of target variable .From the above distplot we can say that the distribution is right skewed there it consists of outliers but we cannot treat these outliers because the rent of the room vary depends upon the several other factors and very high or very low rent is possible.

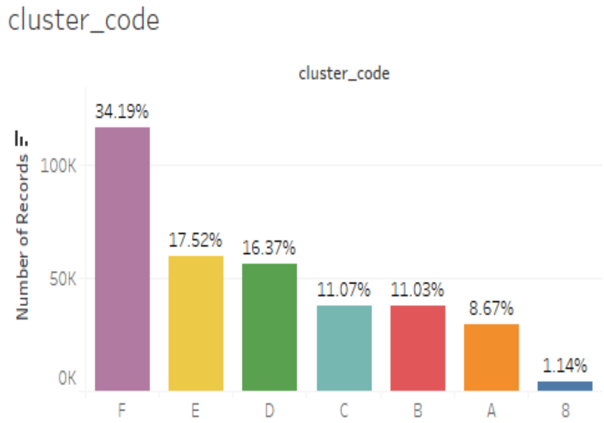
5.2 Univariate Analysis for Categorical features

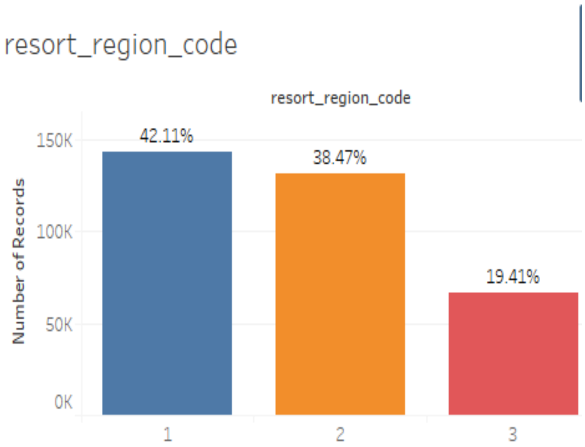


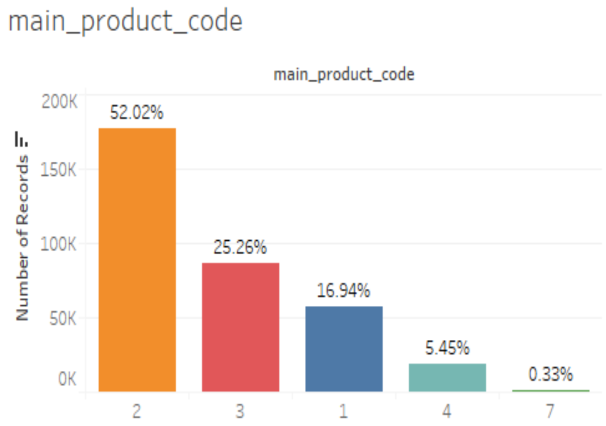


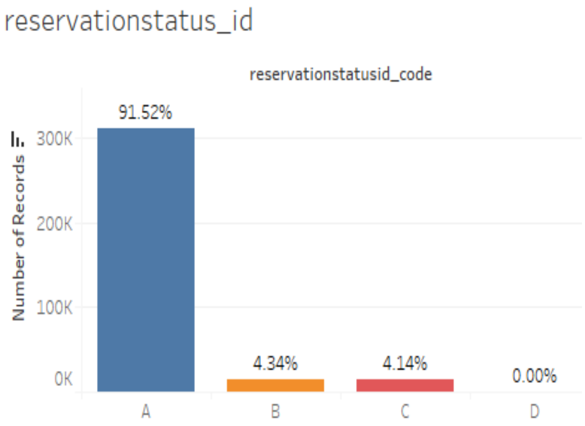




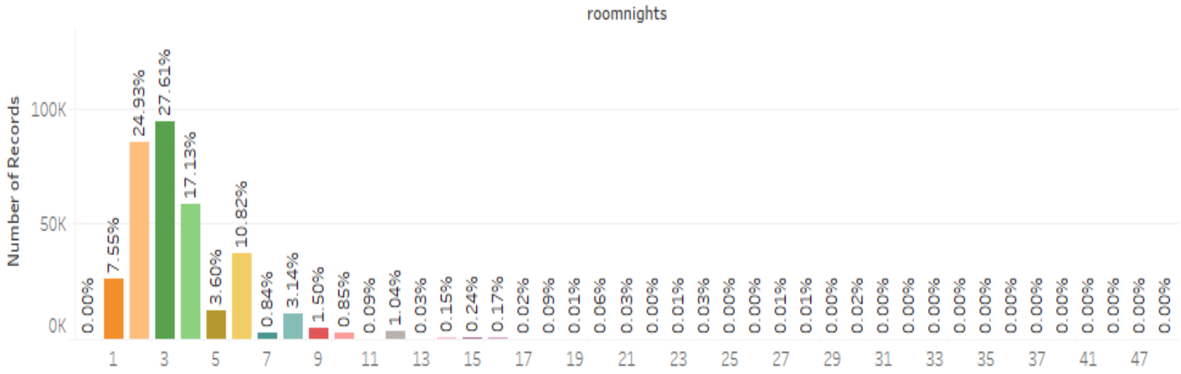




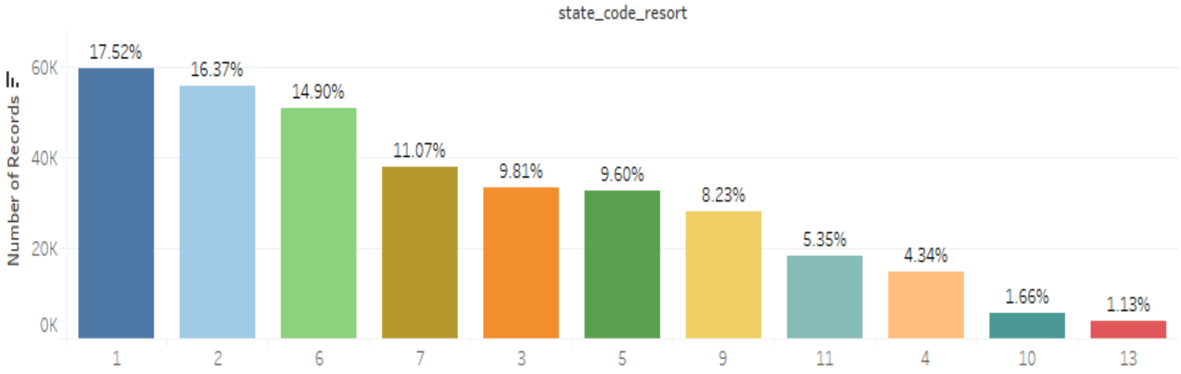




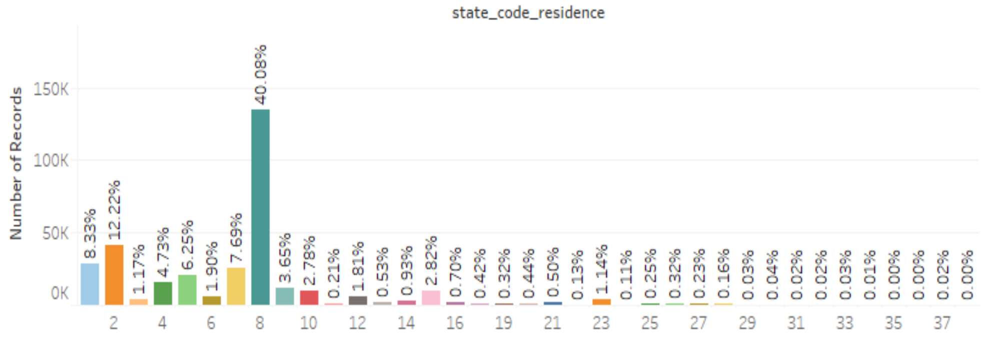
roomnights



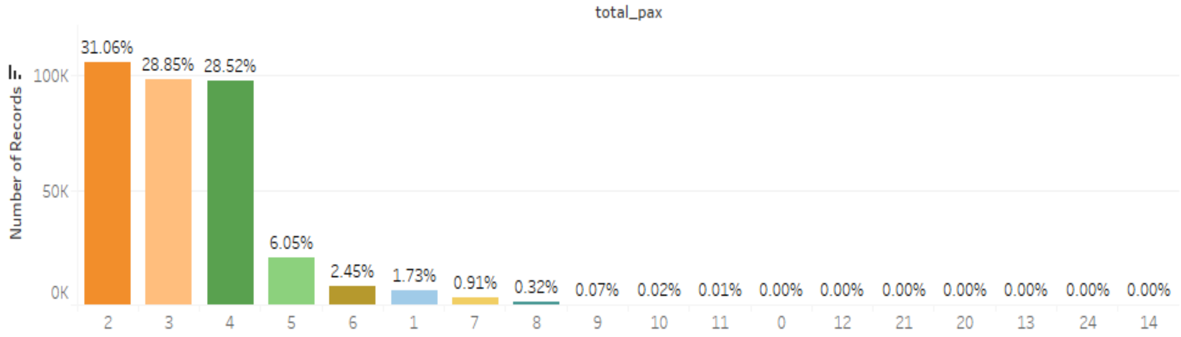
state_code_resort

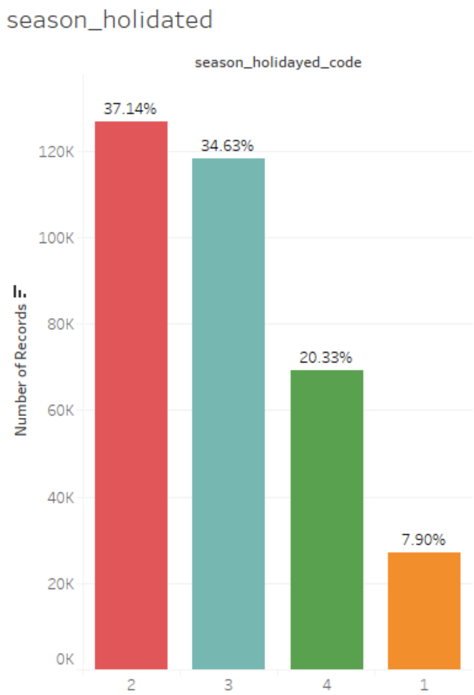


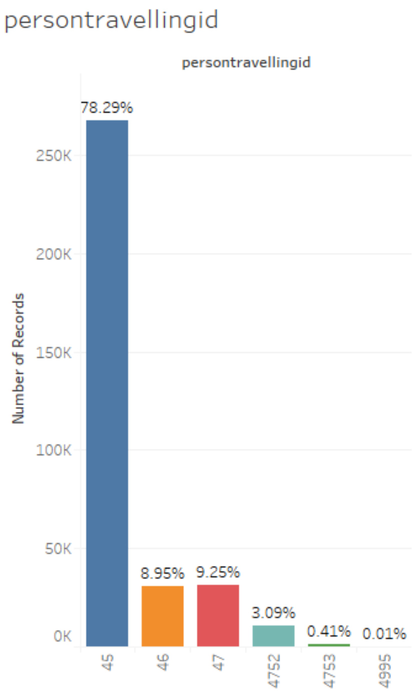
state_code_residence

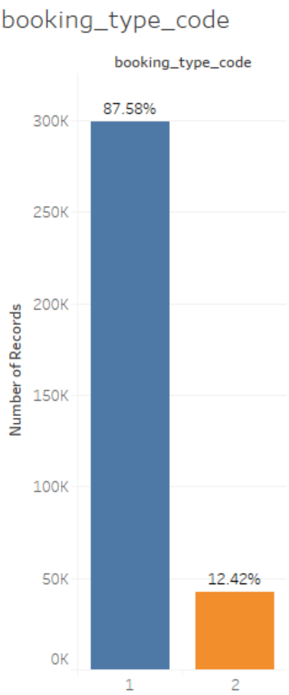


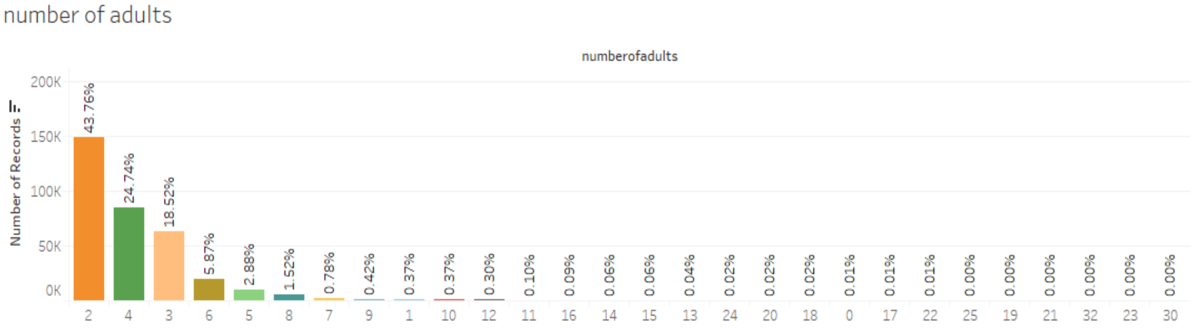
total_pax

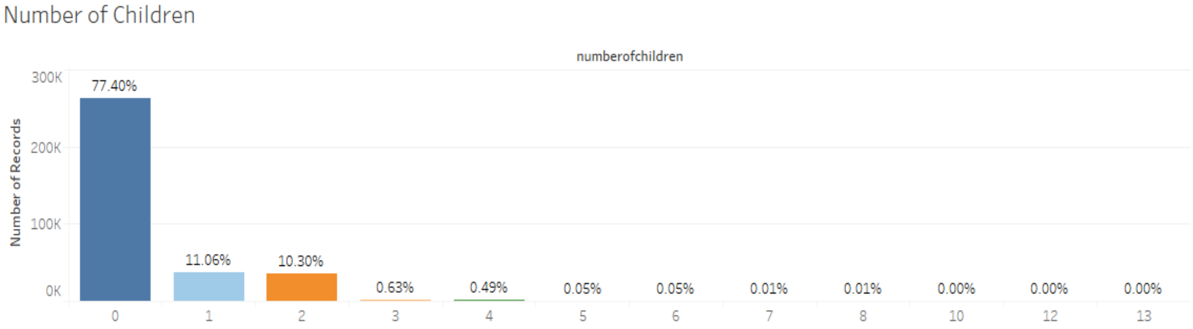










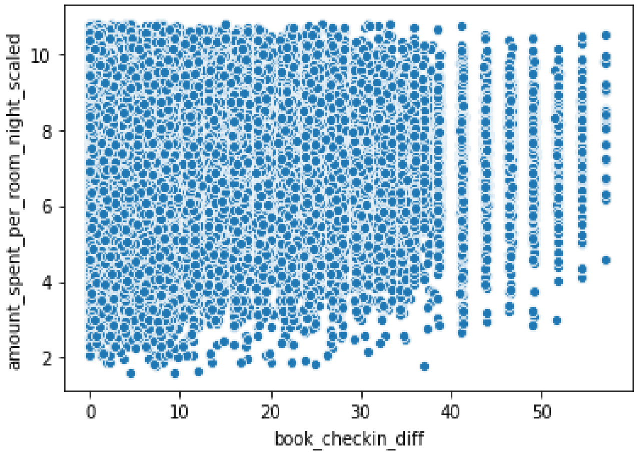


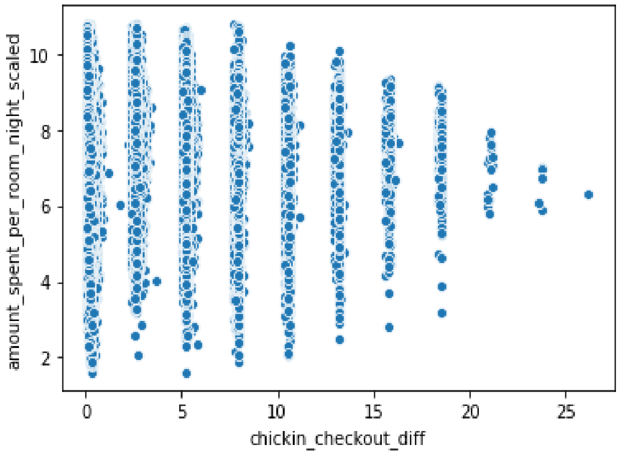
From the plots above we get a fair idea of how many categories each feature contains and the percentage of observation each categories contribute. By analysing the above plots, we can see the nature of customer demands of all these distinct categories within features. For example, by observing numberofchildren we can see that in nearly 77.40% of bookings only adults are the visitors and there are no children with them. By looking at season_holidayed_code we can analyse which season is most preferred for travelling and which is not. By analysing booking_type_code we can say that the type of booking being preferred by the customers etc.

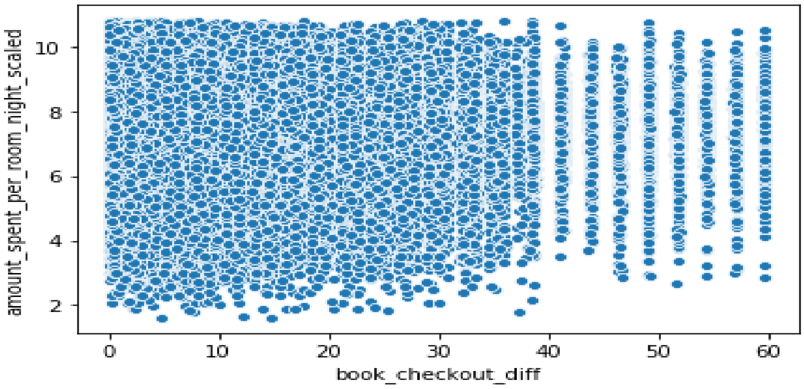
5.3 Bivariate Analysis for continuous features

In this section we are plotting graphs between continuous features just to know the relationship they form with each other which will help us in deciding the better models and will know how much they are correlated with each other.

In these graphs three features are plotted i.e. book_checkin_diff, book_checkout_diff and checkin_checkout_diff against **amout_spent_per_room_night_scaled** (target variable).





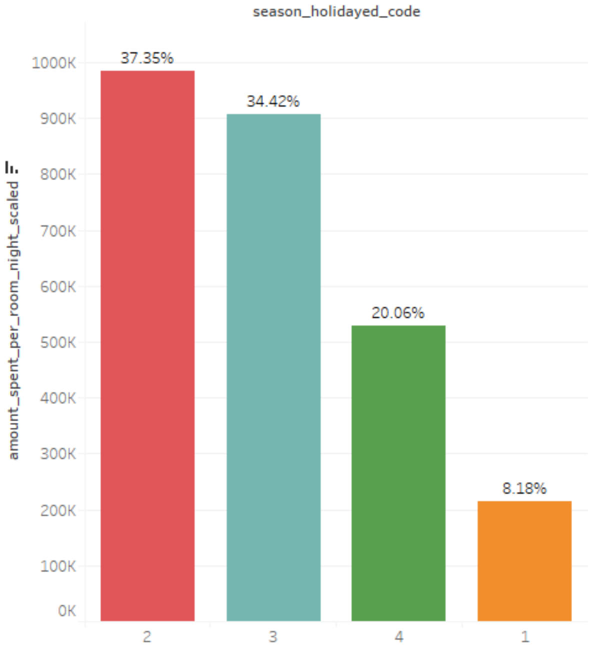


From the above graphs we can conclude that the features are not in linear relationship with the target variable thus linear model is not suitable. We can also conclude that they are not strongly co-related with target variable.

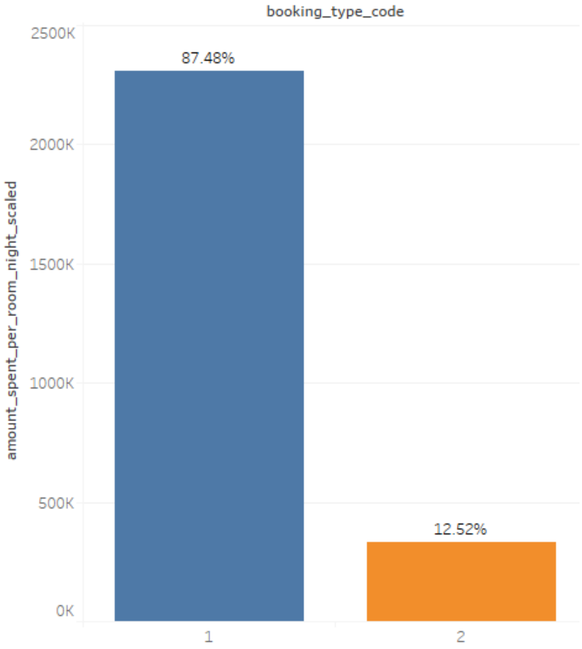
5.4 Bivariate Analysis between Categorical features and Target variables

Here we will draw a plot between all the categorical features and target variable to know how much each category in a feature contributing to the rent of the room. By using this client can make out for which category within a feature is giving more revenue and which is not .To give stimulus to that category to make it attractive for the customers it can avail some offers for that category.

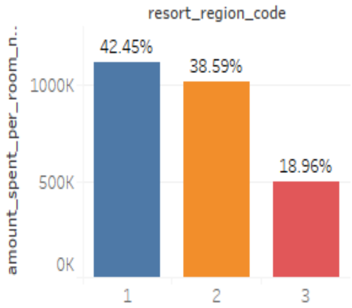
season_holidayed_code vs target variable



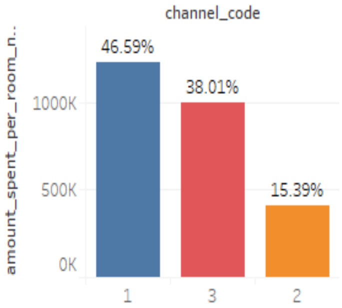
booking_type_code vs target variable



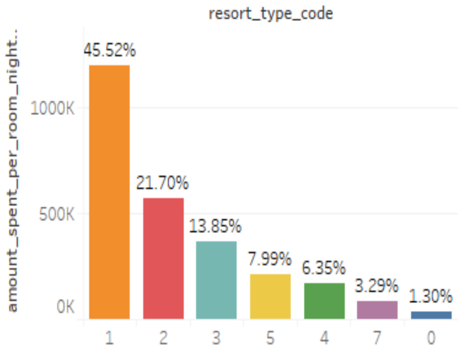
resort_region_code vs target variable



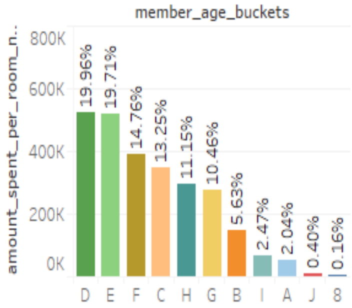
channel_code vs target variable



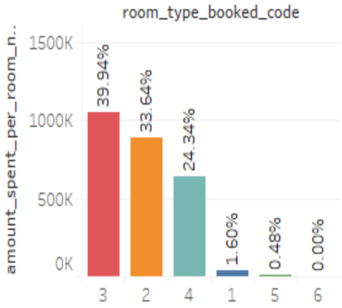
resort_type_code vs target variable



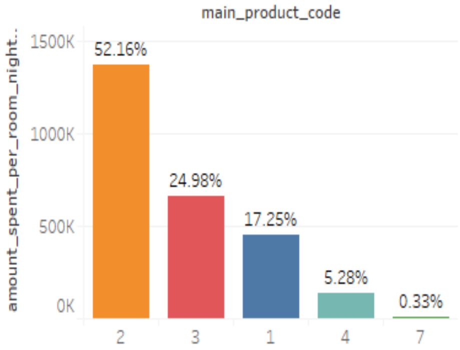
member_age_buckets vs target variables



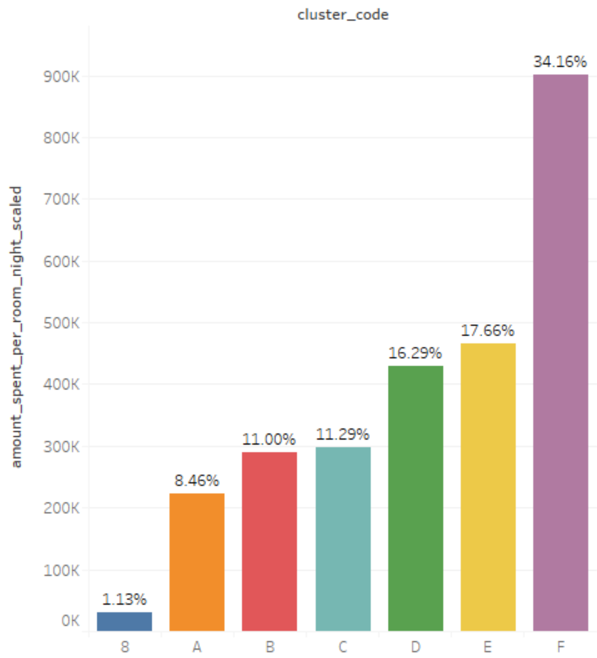
room_type_booked_code vs target variables



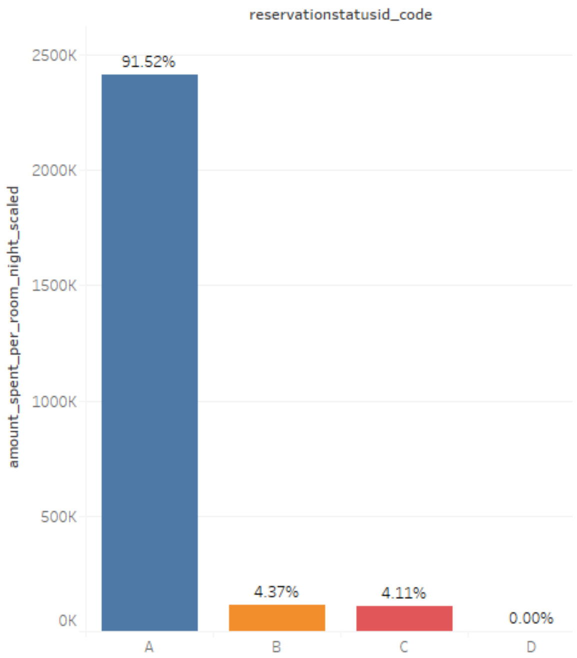
main_product_code vs target variabel



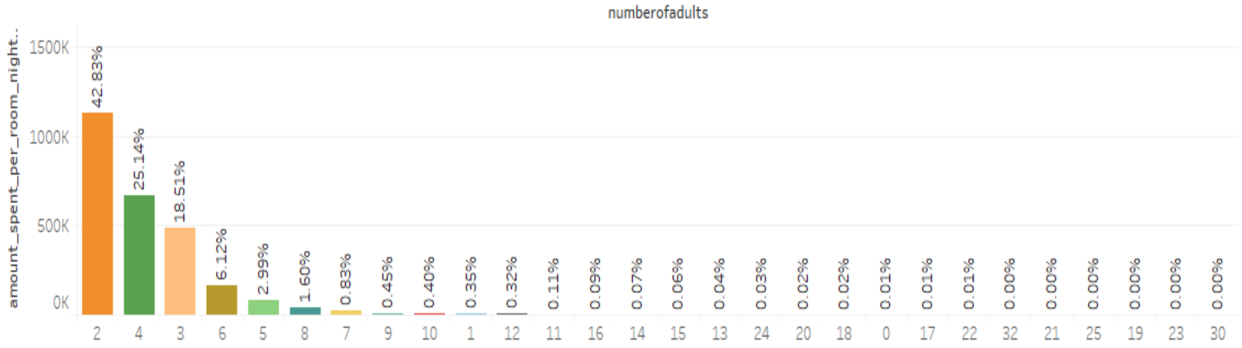
cluster code vs target variable



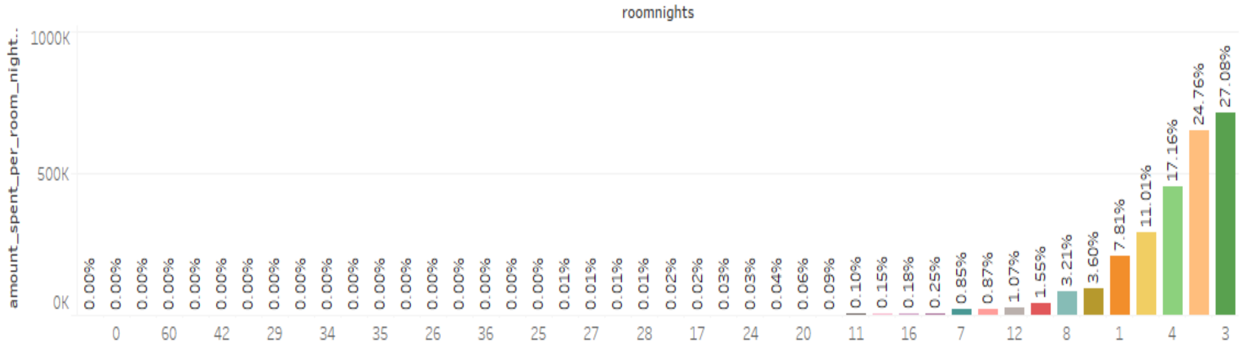
reservation_status_id code vs target variable



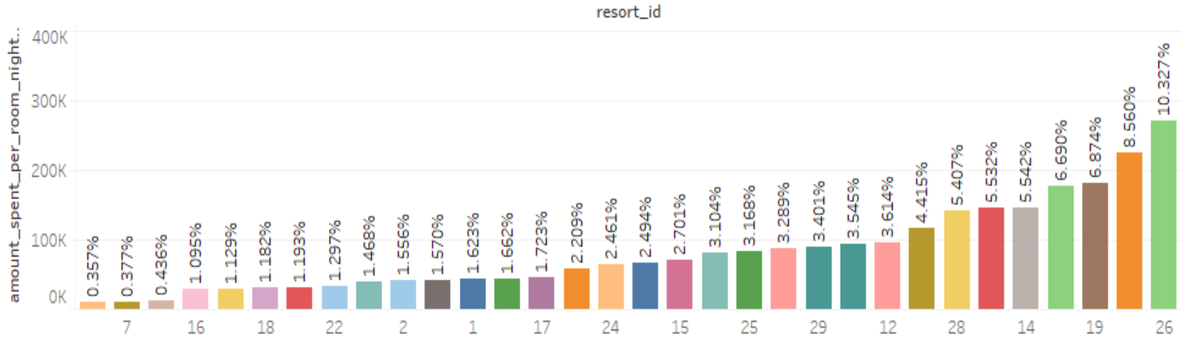
numberofadults vs target variable



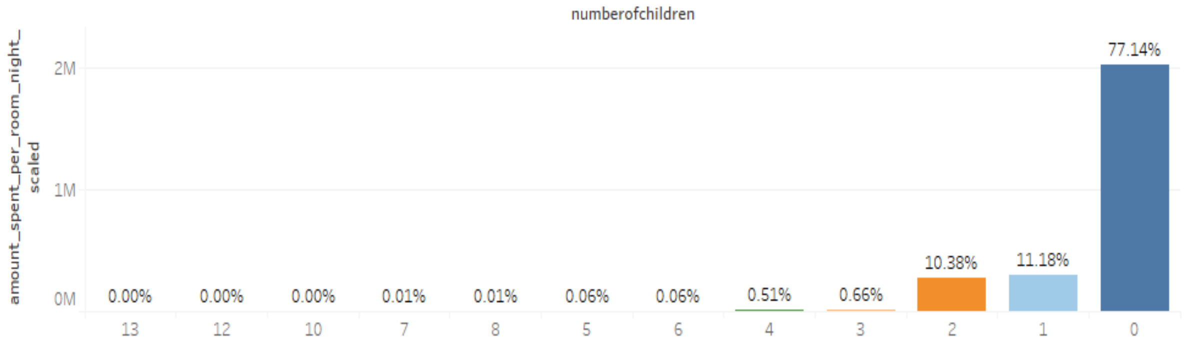
roomnights vs target variable



resort_id vs target variable



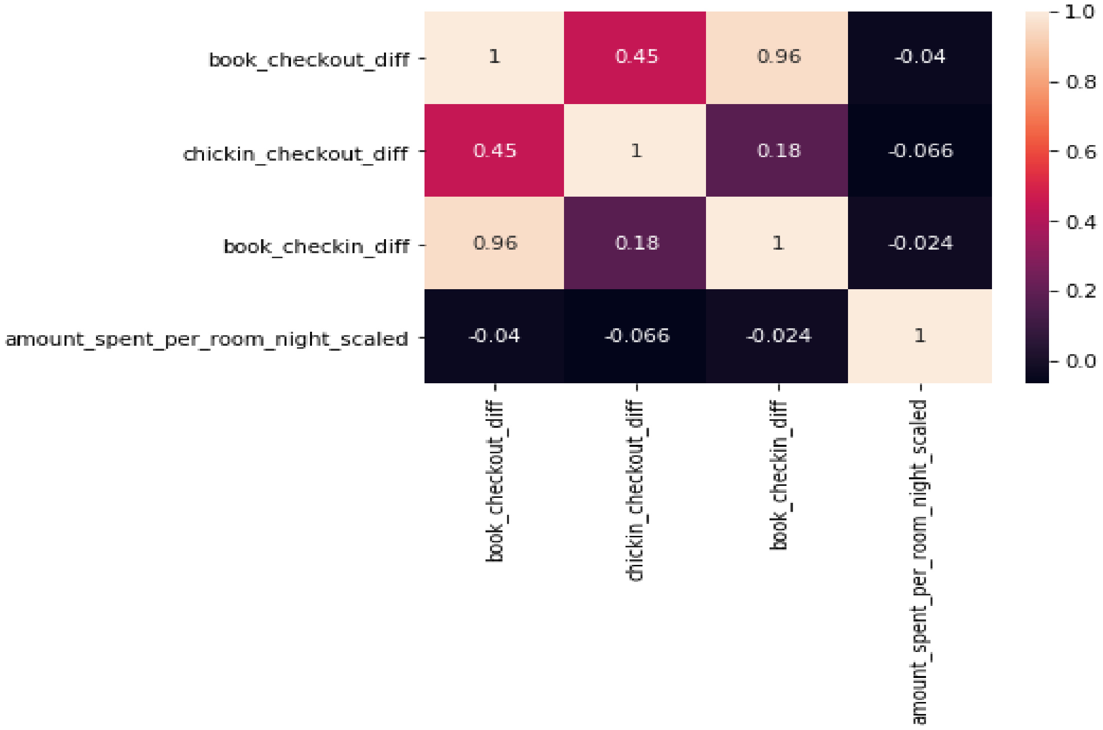
number of children vs target variable



Following points we can conclude from the analysis:

- The average rent paid by the customers of who stay in different region code's is not same so while creating the model we have to consider the feature.
- The average rent paid by the customers who have booked the resort with different channels of bookings is not the same so while creating the model we have to consider the feature.
- The average rent paid by the customers for different number of children is not the same so we cannot dis regard this feature.
- As we can see that there is different demand for booking different type of resort therefore considering resort_type_code is must for our model building.
- We can also see that average rent being paid is different by customers belonging to different age bucket.
- We can see that there is different preference given by different customers as far as type of room is concerned therefore, we have to consider this feature too.
- From the analysis we came to know that few resorts are in high demand in demand as there average rent collection is higher therefore, we have to consider this feature too.
- After analyzing person_travelling_id we can see that for different types of person travelling the rent is different.

5.5 Checking if multicollinearity exists between continuous variables

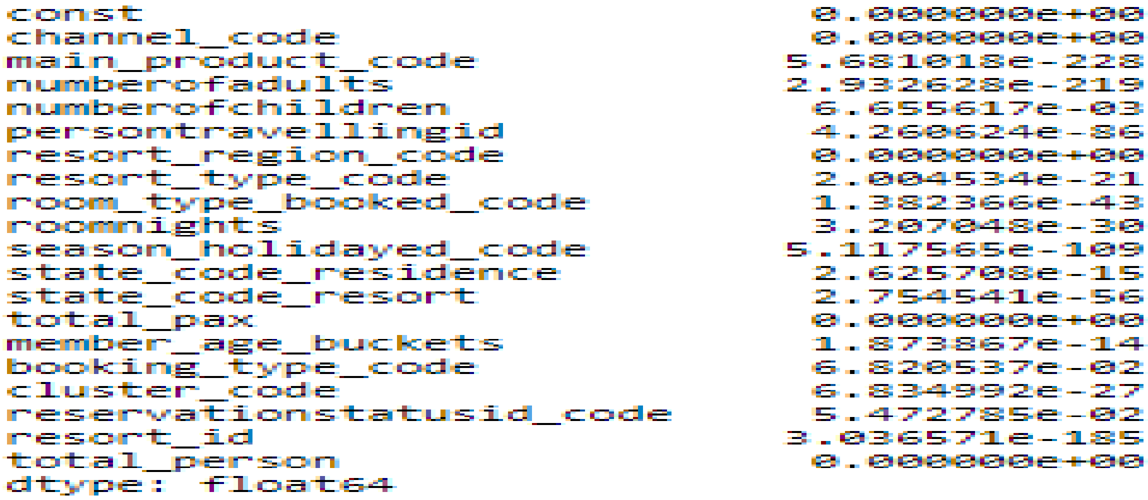


From the above heatmap we can see that there is strong collinearity between **book_checkin_diff** and **book_checkout_diff**. We can see a descent collinearity between **book_checkout_diff and checkin_checkout_diff.** There is very less collinearity between **book_checkin_diff and checkin_checkout_diff.**

5.6 Statistical Analysis

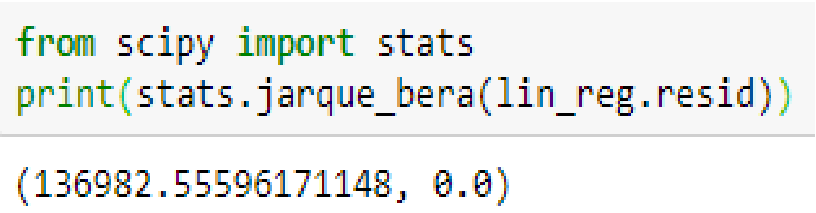
Here we did Anova test between all the categorical variables and target variable. All the features are significant as we can see that all the p_values are lesser than alpha value (0.05).

- **Anova Test Results**



```
const                                      0.000000e+00
channel_code                               0.000000e+00
main_product_code                         5.681018e-228
numberofadults                            2.932628e-219
numberofchildren                           6.655617e-03
persontravellingid                         4.260624e-86
resort_region_code                         0.000000e+00
resort_type_code                           2.004534e-21
room_type_booked_code                      1.382366e-43
roomnights                                 3.207048e-30
season_holidayed_code                     5.117565e-109
state_code_residence                       2.625708e-15
state_code_resort                          2.754541e-56
total_pax                                  0.000000e+00
member_age_buckets                         1.873867e-14
booking_type_code                          6.820537e-02
cluster_code                               6.834992e-27
reservationstatusid_code                   5.472785e-02
resort_id                                 3.036571e-185
total_person                               0.000000e+00
dtype: float64
```

- **jarque_bera test**



```
from scipy import stats
print(stats.jarque_bera(lin_reg.resid))
```

```
(136982.55596171148, 0.0)
```

Here critical value is 137764.8946340364 > 5.99 therefore not residual is not normal.

- **Durbin_Watson Test**

```
Durbin-Watson:                    1.712
```

The value is close to 2. This amount of autocorrelation is tolerable

Part 6 Feature Engineering:

Only three features can be created using available features because except **booking_date**, **checkin_date** and **checkout_date** all the independent features are categorical. Byusing these features **book_checkin_diff** , **book_checkout_diff, checkin_checkout_diff has been created.**

book_checkin_diff is created by subtracting booking_date from checkin_date to know the how early customers are booking the rooms before check_in .

checkin_checkout_diff which has been feature engineered by subtracting by checkin_date from checkout_date to know the how many days customers tend to stay in resort.

book_checkout_diff is created by subtracting booking_date from checkout_date to check if discount is availed to the customer when he is booking earlier and for a long time.

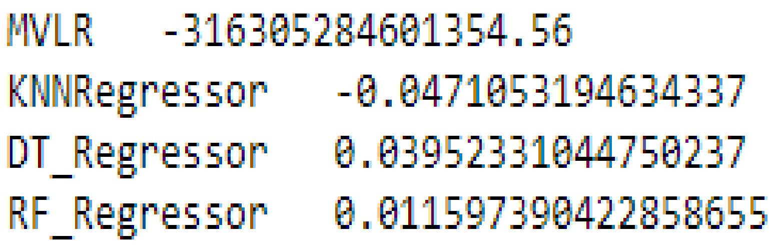
Part 7 Feature Selection

As part of model building ,feature selection has been done to check the number of features which are significant for model building.We used **Backward Elimination method** for feature selection .After using Backward Elimination we are left with **140** features which we will be using for model building.

Part 8 Model Building

This model is the first and most basic model in which all the significant features has been used including three new features which have been created by using date variables.

Note: For regression models R_square metric has been used to measure the performance of the model.



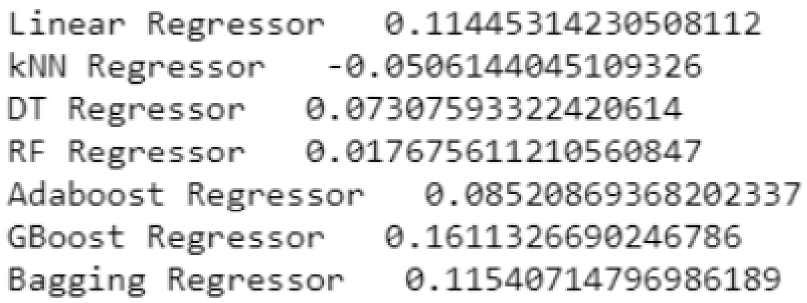
```
MVLR   -316305284601354.56
KNNRegressor   -0.0471053194634337
DT_Regressor   0.03952331044750237
RF_Regressor   0.011597390422858655
```

- Linear Regression Algorithm gives us an under-fit model. This is because no feature is linear in relationship w.r.t to dependent variable (Not following basic assumption of linearity).
- Decision Tree and Random forest are also giving under-fit less variance, more biased). So, here we have to use bagging models from ensemble technique to make the model an generalized model.
- KNN algorithm gives us an under-fit model (less variance, more biased).

<u>Model Score approaches</u>

8.1 First Approach

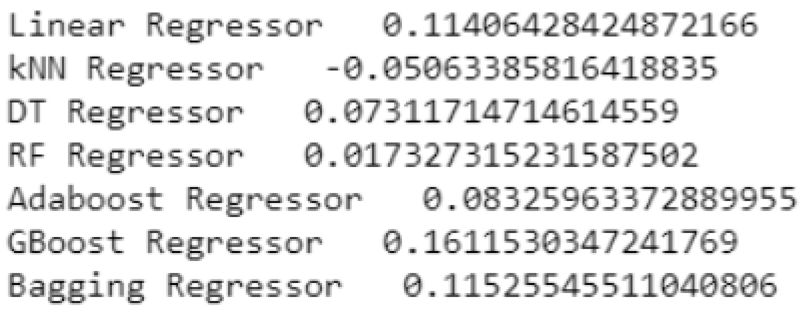
After Clubbing categories within a feature whose value counts are too low.Using this method the number of features after One Hot Encoding decreased.Below are the results of the models.



```
Linear Regressor   0.11445314230508112
kNN Regressor   -0.0506144045109326
DT Regressor   0.07307593322420614
RF Regressor   0.017675611210560847
Adaboost Regressor   0.08520869368202337
GBoost Regressor   0.1611326690246786
Bagging Regressor   0.11540714796986189
```

8.2 Second Approach

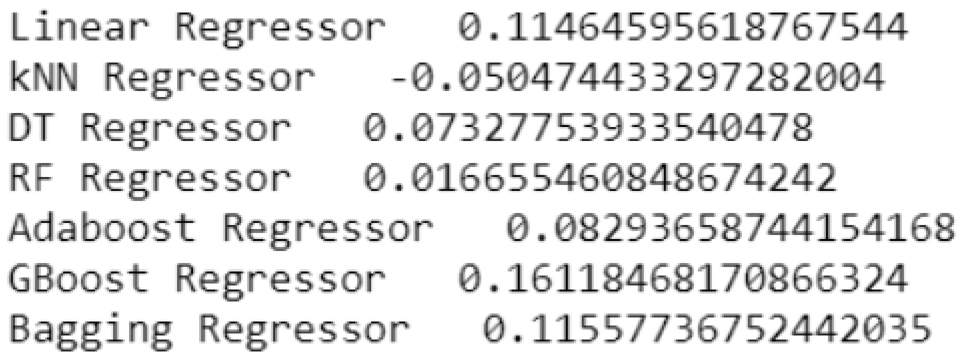
In the second approach we did feature engineering in which we made feature named **start_end_date** .This feature was made by substracing booking_date with new year date and summer holidays start date for years 2015,2016,2017,2018 .If difference between these dates and booking date was between 1 and 4 than in the feature **start_end_date** we put **'1'** and when it is 0 or more than 3 we put the value as **'0'**.In laymen term we made the feature binary on the basis of condition .Below are the scores we got after adding this feature.



```
Linear Regressor   0.11406428424872166
kNN Regressor   -0.05063385816418835
DT Regressor   0.07311714714614559
RF Regressor   0.017327315231587502
Adaboost Regressor   0.08325963372889955
GBoost Regressor   0.1611530347241769
Bagging Regressor   0.11525545511040806
```

8.3 Third Approach

In the third approach we did feature engineering in which we made feature named **quarter_start_end_date** .This feature was made by substracing booking_date with start date and end date of every quarter for years 2015,2016,2017,2018 .If difference between these dates and booking date was between 1 and 4 than in the feature **quarter_start_end_date** we put **'1'** and when it is 0 or more than 3 we put the value as **'0'.**In laymen term we made the feature binary on the basis of condition .Below are the scores we got after adding this feature.

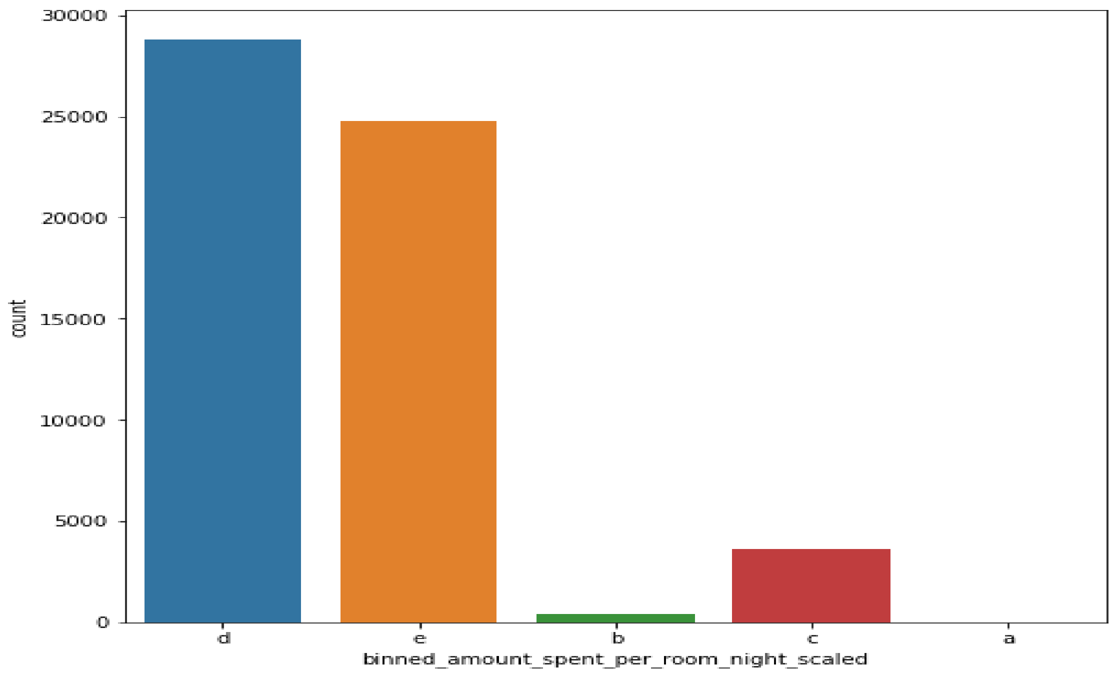


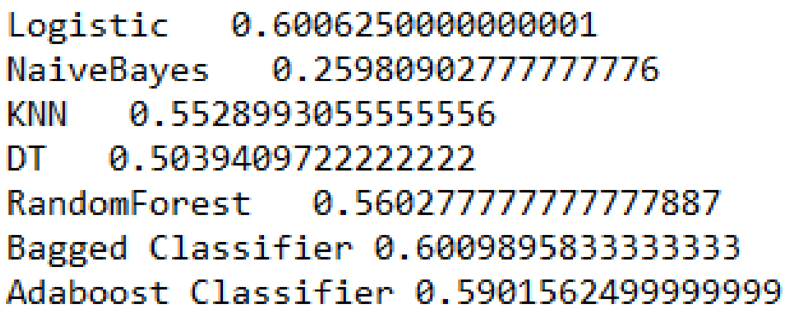
```
Linear Regressor   0.11464595618767544
kNN Regressor   -0.050474433297282004
DT Regressor   0.07327753933540478
RF Regressor   0.016655460848674242
Adaboost Regressor   0.08293658744154168
GBoost Regressor   0.16118468170866324
Bagging Regressor   0.11557736752442035
```

Part 9 Score after changing the nature of prediction from regression to classification

After trying all possible way to increase the score through regression we gave a try to find available clusters inside the dataset but we didn't found distinct clusters within the data set.Therefore we created created 5 classes on the basis target variable. Using distplot we found that the target variable more suitable can be classified into 5 categories named as **'a','b','c','d','e'**.

We didn't used other feature because while doing eda we found that none of the feature can be used for classification because there was no category wise distinction within the feature which was influencing the target variable.The target variable was simply influenced by more demand of each type within the category of the feature .





```
Logistic   0.60062500000000001
NaiveBayes   0.25980902777777776
KNN   0.5528993055555556
DT   0.5039409722222222
RandomForest   0.5602777777777887
Bagged Classifier 0.6009895833333333
Adaboost Classifier 0.5901562499999999
```

These are the **Recall** or **True Positive Rate** for the different models we build .Here we have used Recall metric due to data imbalance.Here we can see using Logistic Regression we have the best score.

Recommendation

The bussiness is expected to avail offers and need to do cost segmentation as per demand .As per the data rent is uniform despite different demand at different region in different season. This argument can be supported by the following factors.

1. There were no distinct clusters found which could have been there if demand of all the features would have been effecting the rent.
2. This can be also supported by our models which we have build for regression.There was no distinctive relation between the no.of days spent in the resort and the rent .Which can be seen from the R_square score of our Linear Regression model.

3. Even our Decision Tree model was not able to find distinct split among the features for enough information gains.Which means that there were no set of conditions which were satisfied which points to the rent charged.

Conclusion

The main objective of this project was to increase operational efficiency through data insights and to predict the rent .As far as data insights are concerned ,there were large number of deformities in this data.After cleaning the data exploratory data analysis is being done.

Using different models and using different approaches we tried to improve the accuracy .In regression we tried to increase the score (**R_Square**) metric using diffenrent approached but the score didn't improve significantly.The highest score in regression we got was for GradientBoost Regression 0.1611 ie : 16 %.

After trying different approaches we changed the nature of prediction by dividing the rent into different class .After changing the nature of prediction from regression to classification we were able to predict with significant higher accuracy .Model with highest **Recall** was Bagging Classifier in which Logistic Regression was used as base model.The Recall score was **0.6009**.